AP/ECE Statistics
Summer Packet 2024

This packet covers material you will need to know as you start your study of Statistics in the fall. The packet includes **4 reading assignments**. After you have completed the reading, there is a ***Google Form*** you must complete. All directions are below:

As you complete each Reading Section:

- Highlight or Underline any concept you are unsure about.
- *Email me* with your questions (see contact info under "Questions" below…)
- Take Notes

  Following each reading section there are instructions on which items you should put in your Statistics notebook. This will give us a head start as we dive into the material we need for the UConn and AP components of our course.
- Once you have completed all reading, fill out the ***Stats Summer Work Google Form*** ***(scan the QR code on the last page to access the Google Form).***

  This Google Form counts as the first homework assignment of the year. You can use your notes as you complete the assignment. Using your resources is highly encouraged!

Due Date:

- The Summer Work Google Form must be completed before the first day of school.
- You will need to bring your Summer Work Packet and your Statistics Notebook with you the first day of school. We will be highlighting key concepts covered in the Summer Work Packet and looking at some applications of these concepts right at the start of school.
- Around the end of the 2nd week of school you will be tested on the material covered in the Summer Work Packet and the ideas we discussed in class connected to these concepts.

Questions?

As you work through the packet, you may have some questions regarding the new concepts. Since many of the ideas are new, that is perfectly normal! To help with this, you can e-mail Mrs. Swanson at mswanson@cpsctg.org, even in the summer! Within a few days I will get back to you!

Reading Section #1

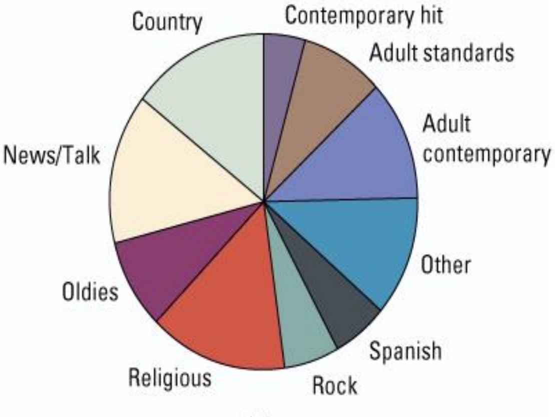
In this first section of Statistics we learn ways to organize data, ways to make pictures of data, ways to describe the overall trend or shape of data, and how to find and use numerical values that describe the center and spread of a set of data.

To work with data, we first need to know the types of data we may encounter. ***Categorical Data*** is data that represents counts or percents of subjects who fit in different categories. This data is not necessarily numeric in nature. For example, the percent of CHS students who prefer different snack foods would represent categorical data. On the other hand, ***Quantitative Data*** is numerical data collected from subjects. For example, the heights of CHS students would represent quantitative data. Our study of ways to organize data will begin with categorical values.

EXAMPLE: The radio audience rating service Arbitron places the country's 13,838 radio stations into categories that describe the kind of programs they broadcast. The table below describes the distribution of formats:

Format	Count of Stations	Percent of Stations
Adult Contemporary	1556	11.2
Adult Standards	1196	8.6
Contemporary Hits	569	4.1
Country	2066	14.9
News/Talk/Information	2179	15.7
Oldies	1060	7.7
Religious	2014	14.6
Rock	869	6.3
Spanish Language	750	5.4
Other Formats	1579	11.4
Total	13838	99.9

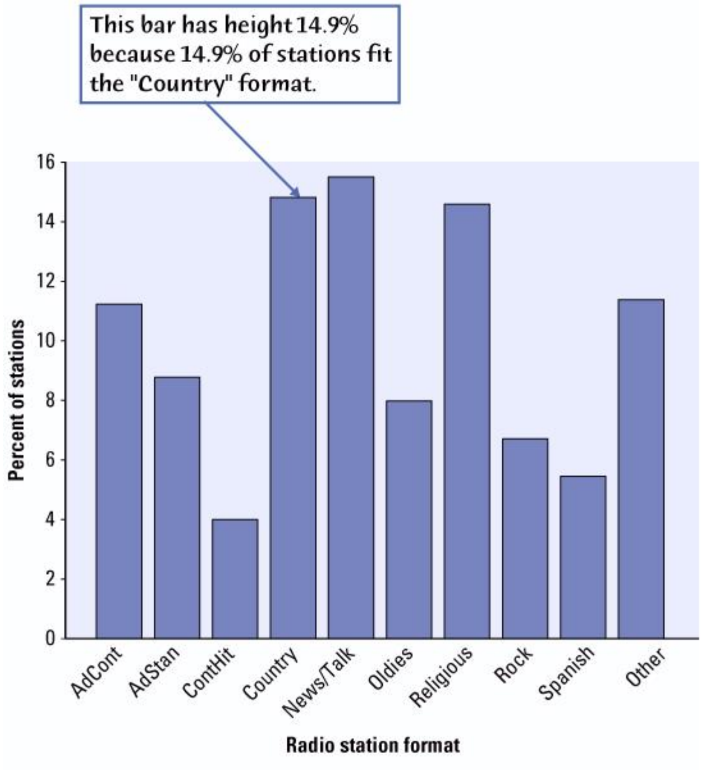
One possibility for organizing this data graphically is to create a ***Pie Chart***, like the one below.



(a)

A pie chart must include all the categories that make up the whole data set. Pie charts are best used when you only want to emphasize each category's relation to the whole, since the graph gives no other information about the data set. Additionally, pie charts can be awkward to make by hand. Generally we will use computer software to generate pie charts in Statistics.

Another possibility for organizing our categorical data is to create a ***Bar Graph***, like the one below.



This bar has height 14.9% because 14.9% of stations fit the "Country" format.

(b)

Notice the bar graph identifies each category in our data set along the horizontal axis, and the vertical axis identifies the percent (or sometimes the count) of data in each category. Additionally, notice the bars in the bar graph are <u>not</u> touching. This is because the categories are not connected. Bar graphs are easier to make and also easier to read than pie charts. Bar graphs allow you to compare quantities in different categories because the bars are measured in the same units.

It is important to notice for any type of graph we make, <u>labels</u> must be included. Without them the reader does not know what data we are considering, what we are trying to compare, or how to draw any conclusions based on the picture presented.

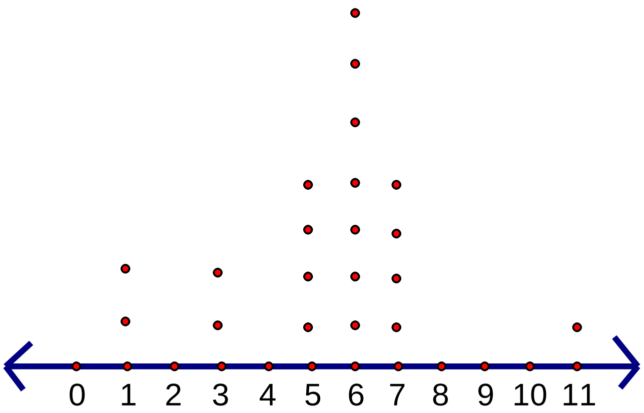
Now let's consider graphs that can be used with quantitative data.

EXAMPLE: A business tracked the number of breakdowns each of their 20 delivery trucks had over the course of a year. The data they collected is below:

6	11	5	1	6	6	7	5	7	6
1	7	5	3	6	3	5	6	7	6

Since this is a small data set, one graph we could make of the data is called a ***Dotplot***. A dotplot for the truck break down data follows:

Truck Breakdown Data



The benefit of this graph is that you get a picture of the data, including which values are more frequent and which are less frequent, while preserving the actual data in the picture. From this example we can see that it was most common for a truck to have between 5 and 7 breakdowns during the year, with 6 being the most frequent value.

The other very common graph we can use to organize quantitative data sets, particularly small quantitative data sets, is the ***Stemplot***, or ***Stem-and-Leaf Plot***. Like the dot plot, the stemplot uses the actual data in the graph, allowing us to see the data and any patterns that exit in the data.

Steps for Creating a Stemplot (or Stem-and-Leaf Plot)

1) Separate each data value (or observation) into a **stem** (consisting of all but the right-most digit, regardless of where any decimal point might lie), and a **leaf** (the final digit of each observation). Stems can be any number of digits, but leaves must be just one digit.
2) Write the stems in a vertical column (smallest to largest) and draw a vertical line to the right of this column of values.
3) Write each leaf in the row to the right of its stem. You can put the leaves in numerical order, but it is not required.

EXAMPLE: Mayor of one region in North Carolina wanted to collect data on the age of homes in their district. The region has two very different types of homes, depending on if they are along the coast (called the Beachside region), or if they are in the hill side (called the Mountain Side region). There is some funding available to help homeowners with older homes make improvements to increase the value of their homes, so the mayor wants to get a picture of where more of the money would go, to Beachside or Mountain Side homes. The table below shows age of a sample of homes in each region.

Age of Beachside Homes									***Age of Mountain Side Homes***								
(in years)									***(in years)***								
22	47	57	62	67	71	78	89	77	85	75	62	45	31	24	25	32	57
70	65	62	54	35	22	49	83	76	64	76	25	78	56	67	26	64	61
68	62	57	53	81	64	70	58	83	32	74	21	37					
27	73																

To begin, let's make a stemplot of JUST the Beachside data. We will use the digits in the tens place as our stems, and the digits in the ones place as our leaves. The stems would look like:

2	
3	
4	
5	
6	
7	
8	

And the completed stemplot would look like:

2	2 2 7
3	5
4	7 9
5	3 4 7 7 8
6	2 2 2 4 5 7 8
7	0 0 1 3 6 7 8
8	1 3 3 9

Key: 2 | 20 = 22 years old

Notice the key for our stemplot. The key is a crucial part of the graph that explains the scale of the data values. From our picture of the data, it appears there are two clusters, suggesting that while many Islamic countries have reasonably high female literacy rates, there are a few that have very low rates comparatively. This indicates a lot of spread in our data.

EXAMPLE: Are the Beachside homes older than the Mountain Side homes, or vice versa? Is their one region that would receive more of the funding for home improvements? To answer this question, we need to compare the ages of homes in the two regions.

To do this, we will make a ***Back-to-Back Stemplot***. For this plot we will put the leaves for the Beachside data on the left, and the leaves for the Mountain Side data on the right. Also, the plot will still need to have a key!

Beachside		*Mountain Side*
7 2 2	2	1 4 5 5 6
5	3	1 2 2 7
9 7	4	5
8 7 7 4 3	5	6 7
8 7 5 4 2 2 2	6	1 2 4 4 7
8 7 6 3 1 0 0	7	4 5 6 8
9 3 3 1	8	5

Key: 2 | 2 = 22 years old

Now we can compare the data sets. Notice that the Beachside data reasonably unimodal with more older homes (between 50 and 70+ years), while Mountain Side data appears to have two clusters, with homes between 20-30 years, and another set between 60-70 years old. Overall, the sample data suggest there are slightly more older homes in the Beachside area, suggesting this region would get a higher percentage of the funds available.

Another consideration to make when using stemplots to describe data is the scale of the data. Scale can influence not only how we set up the key for our data, but also how we choose our stems.

EXAMPLE: The data below represents Reading Scores for a group of 5th graders.

10.1	12.3	10.6	11.7	10.2	12.4	10.8	12.6	10.0	11.5	10.3	11.0
12.2	12.4	11.4	12.8	12.5	11.2	10.9	11.6	11.4	10.9	10.2	10.1
11.2	11.7	12.7	12.2	11.4	11.0	11.3	11.0	11.7	10.5	10.2	11.8
12.7	10.9	11.8	11.3	12.4	10.1						

For this data, the digits to the left of our decimal place will be our stems, and the digit after the decimal will be the leaves. The stemplot will look as follows:

10	0 1 1 1 2 2 2 3 5 6 8 9 9 9
11	0 0 0 2 2 3 3 4 4 4 5 6 7 7 7 8 8
12	2 2 3 4 4 4 5 6 7 7 8

Key: 10 | 0 = 10.0

Notice that the key for this graph tells us where the decimal point lies for each data value. Because of this, we do not need to include any decimal point in the stemplot itself.

This stemplot gives us very little information about the overall shape of the data since it is so clumped together. To solve this problem, we might consider ***Splitting the Stems*** to get a better picture of the data.

How to Split the Stems of a Stemplot (or Stem-and-Leaf Plot)

1) To split the stems, you will create multiple stems of the same number, but divide up the leaves so the graph becomes more spread out and easier to read.

2) The most common way to split the stems is into two pieces, where the leaves 0 – 4 follow the first stem, and the leaves 5 – 9 follow the second stem.

3) You are not required to split the stems into just two pieces, but you must split the stems so there are an equal number of leaf digits that could be assigned to each stem. (For example, splitting into five pieces divides the leaves into 0 – 1, 2 – 3, 4 – 5,
6 – 7, and 8 – 9.)

For our 5th grade Reading Scores data, the stemplot with split stems would look like:

10	0 1 1 1 2 2 2 3
10	5 6 8 9 9 9
11	0 0 0 2 2 3 3 4 4 4
11	5 6 7 7 7 8 8
12	2 2 3 4 4 4
12	5 6 7 7 8

Key: 10 | 0 = 10.0

Now we get a better picture of the data. We can see that the data does not have a lot of spread, and is centered around scores in the low 11's, suggesting 11.0 – 11.4 is the most frequent score range.

One final word of caution about stemplots. In most cases the data we use should be preserved in the graph we make. This will be the expectation for our work. However, in rare instances, printed material may trim the data for the stemplot. This is not a practice we should follow, but one we should be aware of when working with tables created by media and print sources.

EXAMPLE: The data below represents tuition and fees for the 2005-2006 school year at 37 four-year colleges and universities in Virginia. In addition to the data in the table, the state has 23 two-year community colleges that each charged $2135 for the school year.

College	*Tuition & Fees*		*College*	*Tuition & Fees*
Averett	18,430		Patrick Henry	14,645
Bluefield	10,615		Randolph – Macon	22,625
Christendom	14,420		Randolph – Macon Women's	21,740
Christopher Newport	12,626		Richmond	34,850
DeVry	12,710		Roanoke	22,109
Eastern Mennonite	18,220		Saint Paul's	9,420
Emory and Henry	16,690		Shenandoah	19,240
Ferrum	16,870		Sweet Briar	21,080
George Mason	15,816		University of Virginia	22,831
Hampton	14,996		University of Virginia – Wise	14,152
Hampton - Sydney	22,944		Virginia Commonwealth	17,262
Hollins	21,675		Virginia Intermont	15,200
Liberty	13,150		Virginia Military Institute	19,991
Longwood	12,901		Virginia State	11,462
Lynchburg	22,885		Virginia Tech	16,530
Mary Baldwin	19,991		Virginia Union	12,260
Marymount	17,090		Washington and Lee	25,760
Norfolk State	14,837		William and Mary	21,796
Old Dominion	14,688			

A stemplot of this data with split stems (5 stems for each value) was produced:

0	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
0	
0	
0	9
1	0 1
1	2 2 2 2 3
1	4 4 4 4 4 4 5 5
1	6 6 6 7 7
1	8 8 9 9 9
2	1 1 1 1
2	2 2 2 2 2
2	5
2	
2	
3	
3	
3	4

Key: 3 | 4 = $34,000

Notice that the data has all been trimmed, or truncated. That is, only the dollar value in the thousands has been preserved, and the rest of the digits have been removed. While this practice can make it easier to create a stemplot, it does not preserve the data or even accurately round the data. As a result, we should be aware of how to read stemplots that have been created this way, but realize this will not be a regular practice in our own work.

READING SECTION #1 NOTE TAKING:

The following items from this reading must be included in your Statistics notebook:

- Definition of Categorical Data
- Definition of Quantitative Data
- Definition of Pie Chart
- Definition & Example of Bar Graph
- Definition & Example of Dot Plot
- Steps to Creating a Stemplot
- Examples of a Stemplot, a Back-to-Back Stemplot, and a Stemplot with Split Stems

Reading Section #2

READING:

In the first assignment we looked at graphs that can be used for categorical (or qualitative) data (pie charts and bar graphs), and a few graphs that can be used for quantitative data (dotplots and stemplots). In this lesson we are going to consider some additional graphs that are used for quantitative data.

Let's begin with an *EXAMPLE:* The data below represents scores 3rd graders received on a reading assessment.

40	26	39	14	42	18	25	43	46	27	19	47	19	26	35	34
15	44	40	38	31	46	52	25	35	35	33	29	34	41	49	28
52	47	35	48	22	33	41	51	27	14	54	45				

To work with this data we first need to organize it in a ***Frequency Table***. This table is used to organize quantitative data into groups, called ***classes***. For each class the number of data values in the group, called the frequency, is recorded as well as some additional information.

To create our frequency table, we first need to decide how we are going to organize our data into classes. One important rule about creating classes is that we want to have each class have the same number of possible data values in it. For example, if our first group could include the scores 10, 11, 12, 13 and 14, the ***class width*** would be five since that class has five possible data values. When making a frequency table it is important to have the same class width for all groups.

Since the smallest data value in our data set is 14, and the largest value is 54, we know that our classes need to include all of the values between these two numbers. Often, we start the first class at a round number just below the minimum data value, and end the last class at a round number just above the maximum value. For our example, let's use a class width of 5.

Frequency Table

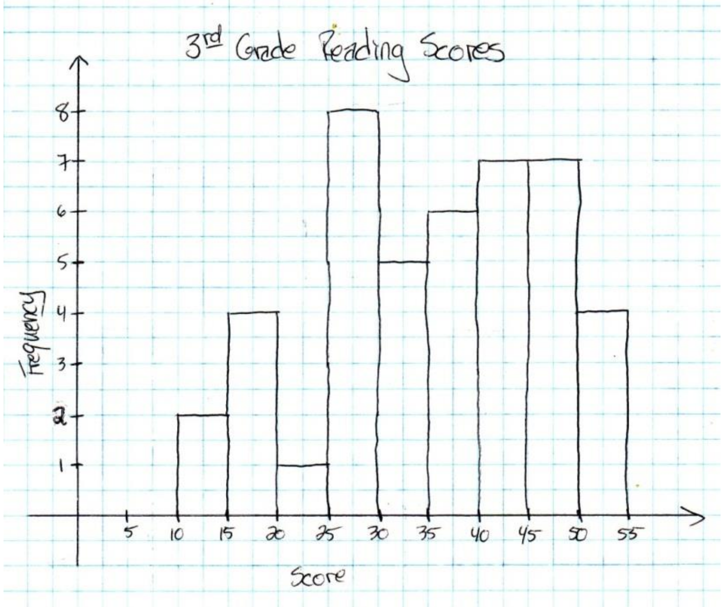
Class	***Frequency***	
$10 \le x < 15$	||	2
$15 \le x < 20$	||||	4
$20 \le x < 25$	|	1
$25 \le x < 30$	𝍸 |||	8
$30 \le x < 35$	𝍸	5
$35 \le x < 40$	𝍸 |	6
$40 \le x < 45$	𝍸 ||	7
$45 \le x < 50$	𝍸 ||	7

50 $\leq$ x < 55	||||	4

Notice that each class includes the data value for the lower end, but not the data value for the upper end. We create our classes this way so that all data values, including any decimal values, are included in just one of the classes. Also notice that the frequency is often recorded as tic marks, and then tallied in the column.

From our frequency table we can create a ***Histogram*** of the data. A histogram is very similar to a bar graph. However, the histogram is used for quantitative data where all possible data values must be accounted for in the classes and in the scale of the x-axis. As a result, the bars in a histogram are touching to demonstrate that all data values have been included in one of the classes.

A ***Frequency Histogram*** of our 3rd grade reading assessment data is below. The graph is called a frequency histogram because the scale of the y-axis is the frequency for each class. Notice the axes are labeled, and the width of each rectangle matches the class width.



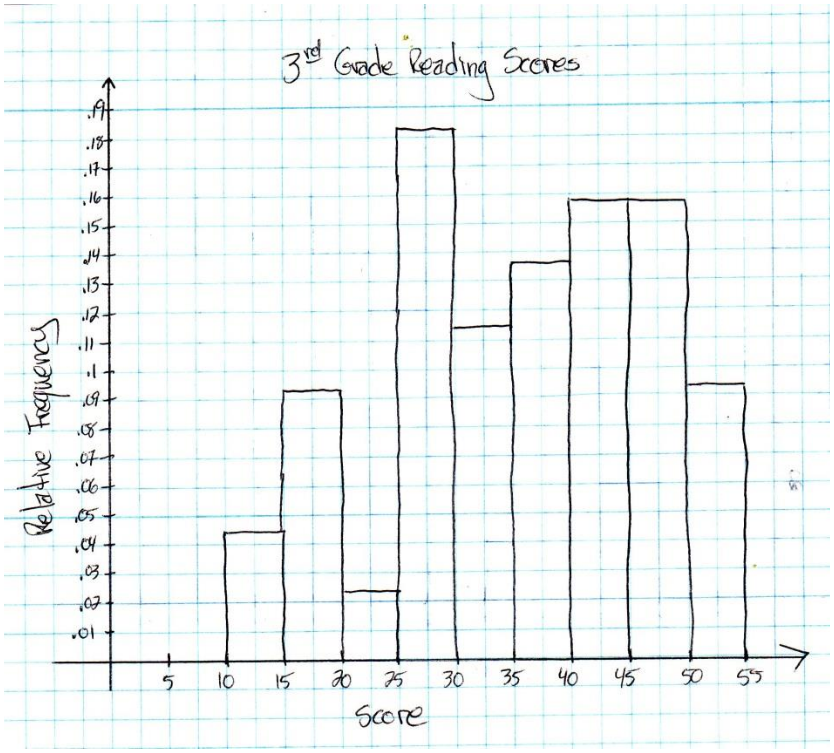
In addition to measuring the frequency for each class, we can also measure the percent of data that falls in each class. This measurement is called the ***Relative Frequency***, and is measured as $r.f. = \frac{frequency}{\#\quad of\quad data\quad values} = \frac{f}{n}$. Notice the relative frequency is the percent in decimal form for each class. Also notice that lower case *n* is a variable commonly used in Statistics to describe the number of data values in a data set, sometimes called the ***Sample Size***.

If we add a column for relative frequency to our frequency table we have:

Class	***Frequency***		***Relative Frequency***
10 $\leq$ x < 15	||	2	$\frac{2}{44} = 0.045$
15 $\leq$ x < 20	||||	4	$\frac{4}{44} = 0.091$
20 $\leq$ x < 25	|	1	$\frac{1}{44} = 0.023$
25 $\leq$ x < 30	𝍸 |||	8	$\frac{8}{44} = 0.182$

$30 \leq x < 35$	卌	5	$\frac{5}{44} = 0.114$
$35 \leq x < 40$	卌 |	6	$\frac{6}{44} = 0.136$
$40 \leq x < 45$	卌 ||	7	$\frac{7}{44} = 0.159$
$45 \leq x < 50$	卌 ||	7	$\frac{7}{44} = 0.159$
$50 \leq x < 55$	||||	4	$\frac{4}{44} = 0.091$

And now we can make a new histogram, called a ***Relative Frequency Histogram*** where the y-axis is now relative frequencies:

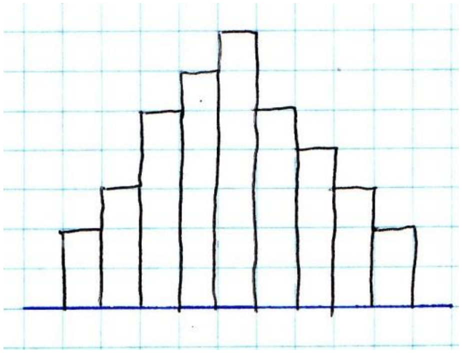


Notice that the overall shape of the graph has not changed from the frequency histogram. The only difference in these graphs is the scale of the y-axis.

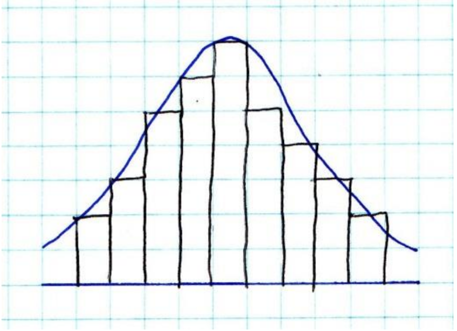
Now that we have considered several ways of organizing and graphing data, we need to consider what these pictures can tell us about a set of data. Often in statistics we are asked to <u>describe</u> the data based on the picture. When we describe the shape of a data set from a graphical representation, this is called ***describing the distribution***. There are three components that <u>must</u> be part of any description: ***shape***, ***center*** and ***spread***. We will begin with the tools used to describe shape.

To develop the terms used to describe the shape of a set of data, consider the diagrams below:

1) When a histogram looks as follows:

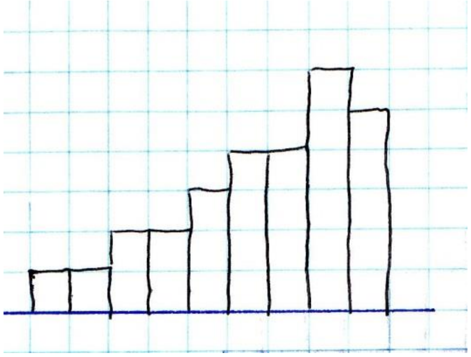


We can think of "smoothing out" the bars to form a curve that looks like:

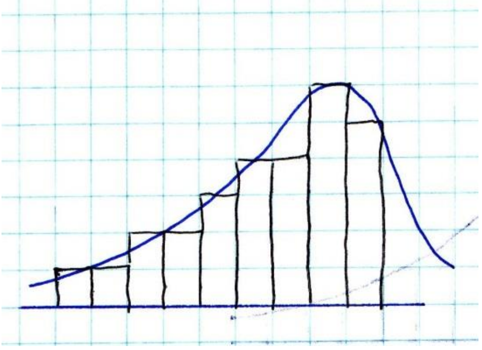


This data is considered to be ***symmetric***, or ***bell-shaped***.

2) When a histogram looks as follows:

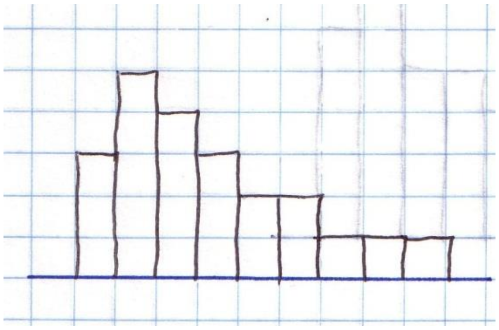


we can think of "smoothing out" the bars to form a curve that looks like:

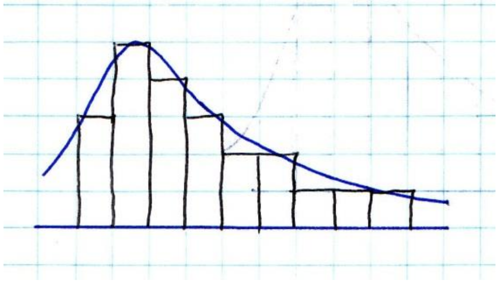


This data is more spread out to the left (or toward the smaller values) since the curve has a longer tail to the left. So, we call this data ***skewed left***.

3) When a histogram looks as follows:



We can think of "smoothing out" the bars to form a curve that looks like:

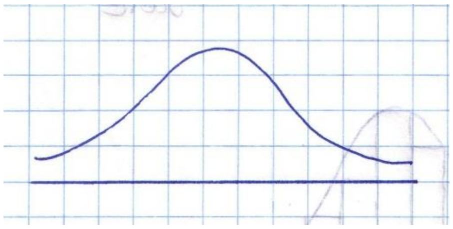


This data is more spread out to the right (or toward the larger values) since the curve has a longer tail to the right. So, we call this data ***skewed right***.

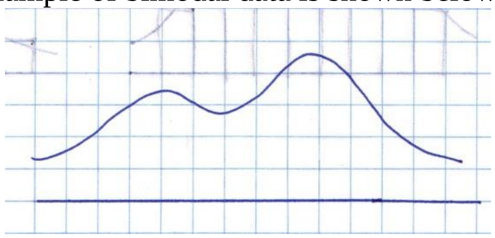
Now, when we describe the ***center*** of a distribution, we are looking for an approximation of where the middle data value lies. As we continue our study of Statistics, we will develop specific calculated values that will describe center for us. However, from the graph of a set of data we also want to consider whether the center of the data seems to be clustered in one area, or perhaps in two different clusters.

Consider these diagrams:

1) When the center of a set of data is clustered in one area, we describe the data as being ***unimodal***. An example of unimodal data is shown below:



2) When the data appears to be clustered in two different areas in a distribution, we describe the data as being ***bimodal***. An example of bimodal data is shown below:

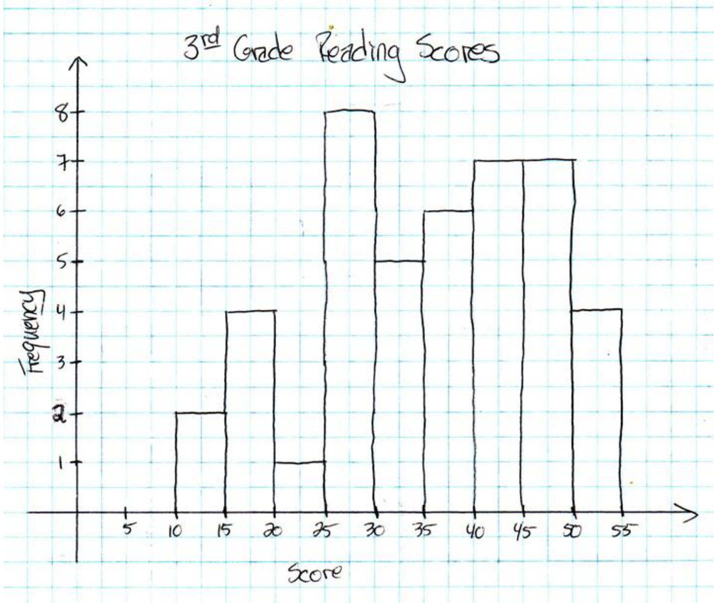


Finally, our description of data should include ***spread***. Here we are trying to describe how far the data is spread out, and if there appear to be any data values that are noticeably different, or far away, from the other data values. As with center, we will continue to develop measures that we can calculate to describe the spread of a set of data. However, the first measure we use as we begin to develop our statistical tools is the ***Range***. The range of a set of data is calculated as the largest value – the smallest value. It is simply a measure of how far along an axis the data spreads.

One other idea we should introduce at this time is the concept of ***Outlier***. An outlier is a data value which appears to be significantly different from the other values in a data set. Outliers can significantly increase the value of the range, and the overall spread of a set of data. While in other disciplines we may want to "throw out" outliers, from a Statistics stand point we never want to ignore an outlier. Perhaps the outlier occurred simply due to measurement error, or perhaps it is an important finding that means something about the data we have collected.

As we continue our work with Statistics, we will develop specific calculations that can help us determine if a value is far enough from the rest of the data set to be considered an outlier. For now, outliers will be quite apparent in our data set and graphs.

So, when we describe the distribution of data, we need to include all of the key components of shape, center and spread. To show you how a complete description might look, consider our histogram from the 3rd grade reading data again:



Remember that when we describe the shape, center and spread of a distribution we think of "smoothing out" the bars of the histogram (or the leaves of a stemplot) to get an overall picture of the data.

From this picture we would describe this data set as follows: "*The data is unimodal and slightly skewed left. The data is centered at approximately 35 – 40 points, and has a fairly large spread with a range of 40 points. There does not appear to be any outliers.*"

Before we wrap up our work in this lesson, we have one more type of graph to consider. This graph, called an ***Ogive*** creates a picture of cumulative frequencies. To see how this works, consider our 3rd grade reading assessment data again. This time we will add 2 more columns to our frequency table. The first column will be ***cumulative frequencies***. These are a tally of how many data values have been accounted for up to the end of that class. So for example, at the end of the class 10 ≤ x <15 we have accounted for 2 of the data values, but by the end of the 15 ≤ x < 20 class we have accounted for a total of 6 data values since the data set has 6 values that are less than 20:

Frequency Table

Class	***Frequency***	***Relative Frequency***	***Cumulative Frequency***	
$10 \le x < 15$	𝍷𝍷 2	$\frac{2}{44} = 0.045$	2	
$15 \le x < 20$	𝍷𝍷𝍷𝍷 4	$\frac{4}{44} = 0.091$	6	
$20 \le x < 25$	𝍷 1	$\frac{1}{44} = 0.023$	7	
$25 \le x < 30$	𝍸𝍷𝍷𝍷 8	$\frac{8}{44} = 0.182$	15	
$30 \le x < 35$	𝍸 5	$\frac{5}{44} = 0.114$	20	
$35 \le x < 40$	𝍸𝍷 6	$\frac{6}{44} = 0.136$	26	
$40 \le x < 45$	𝍸𝍷𝍷 7	$\frac{7}{44} = 0.159$	33	
$45 \le x < 50$	𝍸𝍷𝍷 7	$\frac{7}{44} = 0.159$	40	
$50 \le x < 55$	𝍷𝍷𝍷𝍷 4	$\frac{4}{44} = 0.091$	44	

The last column we add to the frequency table is for ***Relative Cumulative Frequency***. This measure, like relative frequency, is the percent (in decimal form) of the total data values in or below that class. The relative cumulative frequency is calculated as: $\frac{cumulative \quad frequency}{\# \quad of \quad data \quad values} = \frac{cf}{n}$.

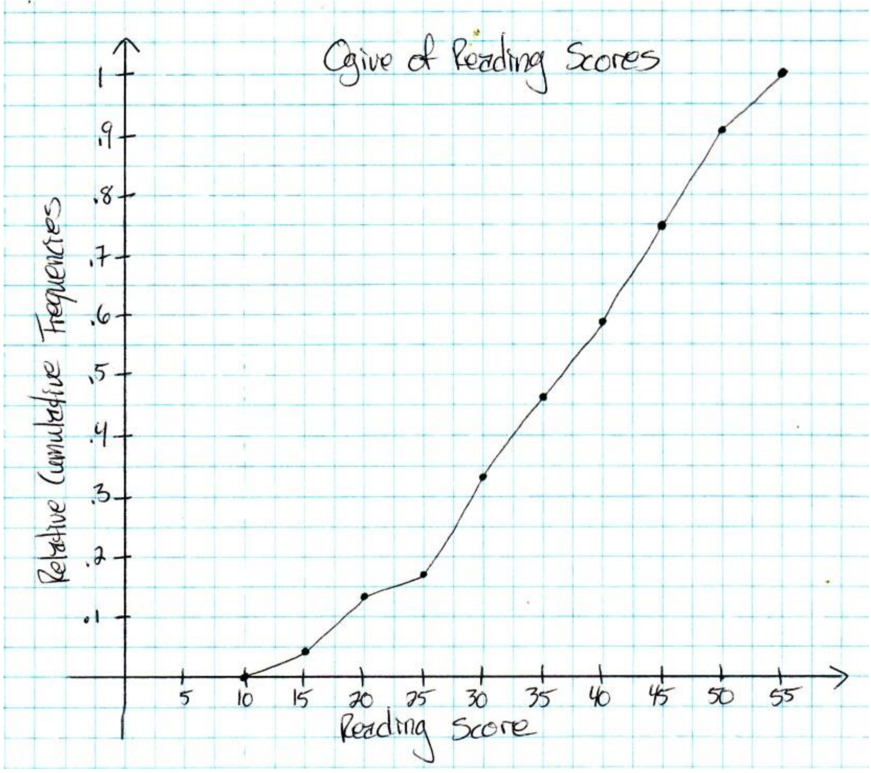
So, our completed frequency table looks like:

(On next page…)

Frequency Table

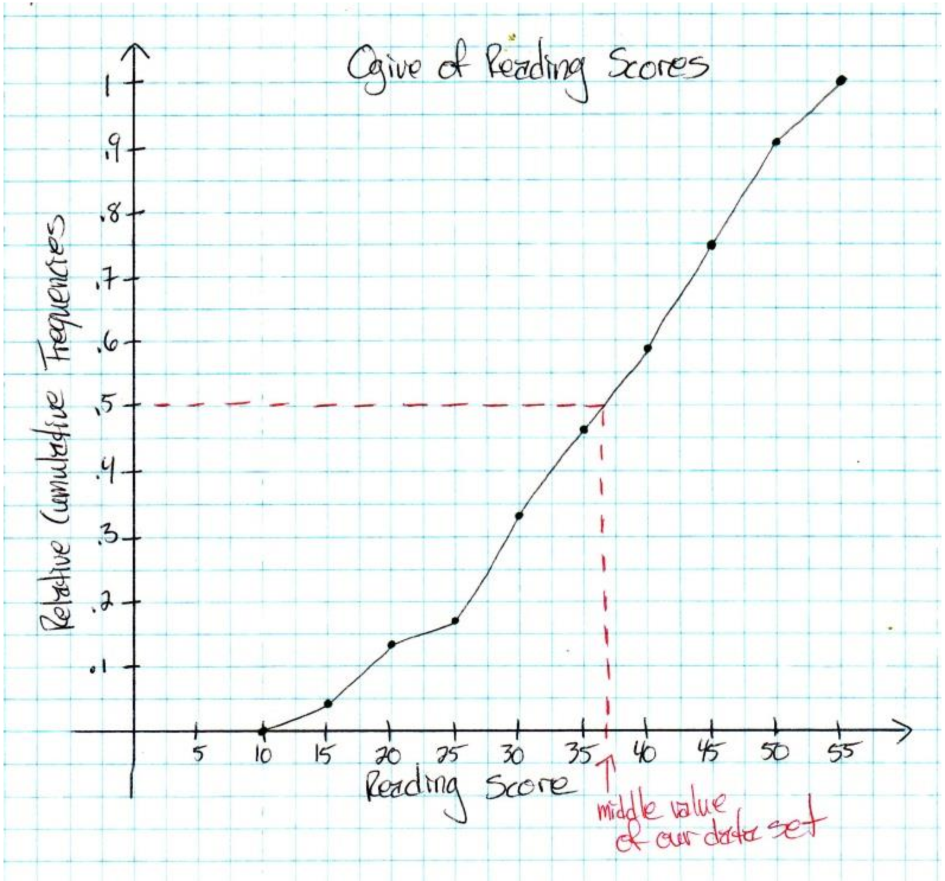
Class	***Frequency***		***Relative Frequency***	***Cumulative Frequency***	***Relative Cumulative Frequency***
$10 \leq x < 15$	||	2	$\frac{2}{44} = 0.045$	2	$\frac{2}{44} = 0.045$
$15 \leq x < 20$	||||	4	$\frac{4}{44} = 0.091$	6	$\frac{6}{44} = 0.136$
$20 \leq x < 25$	|	1	$\frac{1}{44} = 0.023$	7	$\frac{7}{44} = 0.159$
$25 \leq x < 30$	𝍸 |||	8	$\frac{8}{44} = 0.182$	15	$\frac{15}{44} = 0.341$
$30 \leq x < 35$	𝍸	5	$\frac{5}{44} = 0.114$	20	$\frac{20}{44} = 0.455$
$35 \leq x < 40$	𝍸 |	6	$\frac{6}{44} = 0.136$	26	$\frac{26}{44} = 0.591$
$40 \leq x < 45$	𝍸 ||	7	$\frac{7}{44} = 0.159$	33	$\frac{33}{44} = 0.75$
$45 \leq x < 50$	𝍸 ||	7	$\frac{7}{44} = 0.159$	40	$\frac{40}{44} = 0.909$
$50 \leq x < 55$	||||	4	$\frac{4}{44} = 0.091$	44	$\frac{44}{44} = 1$

The relative cumulative frequencies are used to create a graph called an ***Ogive***. This graph plots points that represent the end of each class as the *x* coordinate, and the relative cumulative frequency for the *y* coordinate. The ogive for our 3rd grade reading data is below:



Notice the ogive is a line graph where the points are connected by line segments. This graph can tell is a great deal about our set of data. For *EXAMPLE:* There is a very steep line between the points 25 and 30, suggesting a lot of data in the class $25 \leq x < 30$. In fact, that class has the largest frequency.

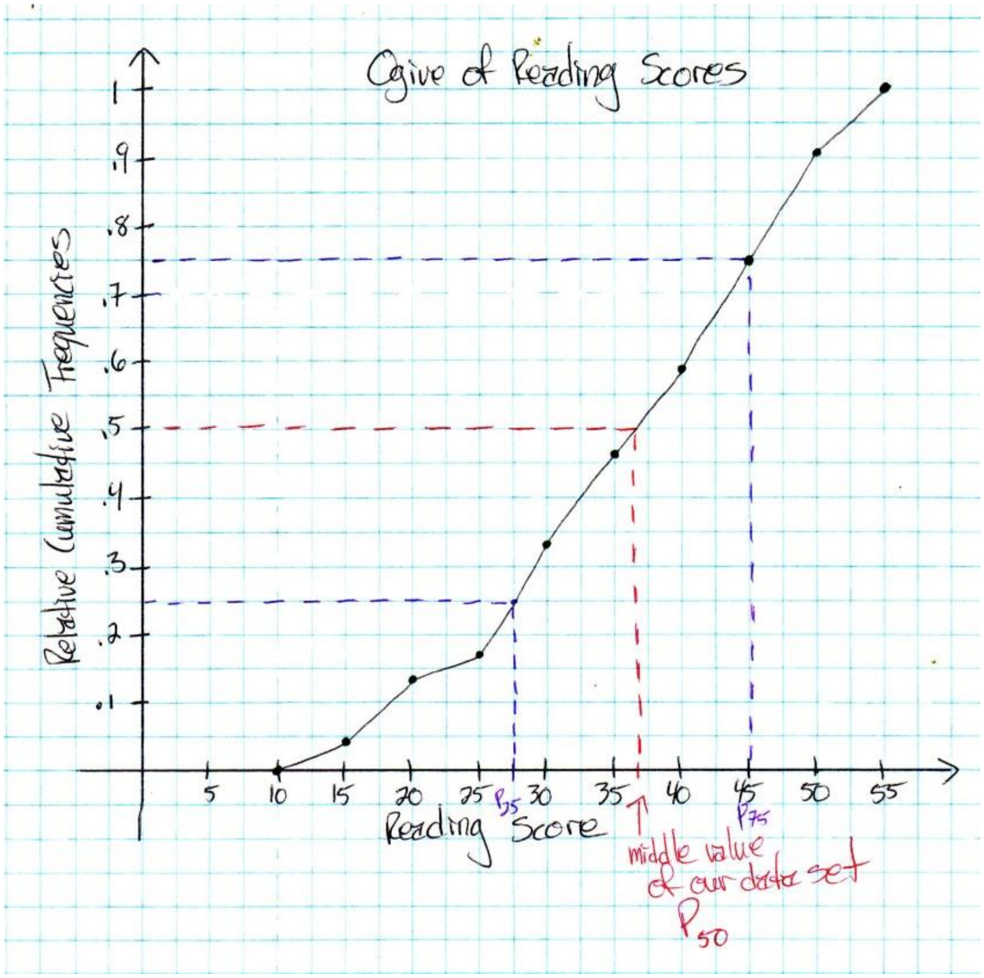
The ogive can also help us identify other information about a data set. For *EXAMPLE:* Where would the middle of our data set be? That is, what is the middle score? To answer this question, we could draw a horizontal line from the relative cumulative frequency value of 0.5. This value divides the data in half since 50% of the data is below this value, and 50% of the data is above this value.



From this graph our middle value is in the class $35 \leq x < 40$. We do not know exactly what the middle value is from the table, but we know where to look to find it.

Connected with the idea of finding where the data is divided into two equal parts (50% below and 50% above), is the idea of ***Percentile***. The ***Percentile*** assigned to any data value is the percent of values that are less than this value. So, for example, if you scored in the 90th percentile on your SATs, you scored better than 90% of students who took the SAT in the same calendar year as you, and your score would be denoted P_{90}.

The idea of percentile is often used to describe data in an ogive. Often, we would like to know the 50th percentile value (P_{50}), or the middle number, as well as the 25th and 75th percentile values (P_{25} and P_{75}). Let's look at our ogive of the 3rd grade reading data again:



From this graph we see $P_{25} \approx 27.5$, P_{50} is in $35 \leq x < 40$, and $P_{75} \approx 45$. Based on these values and our ogive, it does appear that the data is evenly spread between the classes. That is, the quartiles appear to divide the data so there is, relatively, the same number of classes in each quarter. When data has significant skew this is not the case. Instead, one of the quarters (the highest or the lowest) may not appear to have the same number of classes included.

Although we can calculate a variety of percentiles for a set of data, in Statistics we generally look at the 25th, 50th and 75th percentiles that divide the data sent into quarters.

READING SECTION #2 NOTE TAKING:

The following items from this reading must be included in your Statistics notebook:

- Definition & Example of Frequency Table
- Definition of Class Width
- Definition & Example of Histogram
- Definition & Equation for Relative Frequency
- Definition & Symbol for Sample Size
- Description of the difference between a Frequency Histogram and a Relative Frequency Histogram, and an Example of a Relative Frequency Histogram
- Description of how we Describe a Distribution, and the <u>three</u> components that must be included in any description of data
- Definition & Example of Symmetric Data
- Definition & Example of Skewed Left Data

- Definition & Example of Skewed Right Data
- Definition & Example of Unimodal Data
- Definition & Example of Bimodal Data
- Definition of Range
- Definition of Outlier
- Definition of Cumulative Frequency
- Definition of Relative Cumulative Frequency
- Definition & Example of Ogive
- Definition of Percentile

Reading Section #3

READING:

In the last lesson we learned the required elements for describing the distribution of a set of data: shape, center and spread. We also learned specific terminology used in describing shape of the distribution. In this lesson we will develop the statistical values used to describe center and spread of a set of data. We will begin with measures of center.

Measures of Center

1) MEAN – The ***Mean*** of a set of data is the numerical average of the data values. The symbol used for the mean is $\bar{x}$ (read "x bar"), and the formula for the mean is:

$$\bar{x} = \frac{Sum\ \ of\ \ the\ \ Data\ \ Values}{\#\ \ of\ \ Values} = \frac{\sum x}{n}$$

Notice the symbol $\sum$ in the formula. This symbol means to sum the values that follow the symbol. So, $\sum x$ means to add up all the x-values, or individual data values. Also notice the variable n represents the total number of data values.

2) MEDIAN – The ***Median*** is the middle number when a set of data is ordered from least to greatest. If there are an odd number of values, the median is the exact middle number. If there are an even number of values, the median is the average of the two middle numbers. The symbol used for the median is x (read "x tilda").

3) MODE – The ***Mode*** is the most frequently occurring number in a set of data. If more than one value occurs most often, there can be more than one mode for a data set. However, if all data values occur the same number of times, the data set has no mode.

In general, the mode is not a widely used measure of center in Statistics. Instead, the mode is generally used as part of our description of shape (unimodal vs. bimodal).

Now consider this *EXAMPLE:* The table below shows the Highway Fuel Economy (in mpg) of two-seater sports cars produced in 2004.

Model	***MPG***
Acura NSX	24
Audi TT Roadster	28
BMW Z4 Roadster	28
Cadillac XLR	25
Chevy Corvette	25
Dodge Viper	20
Ferrari 360 Modena	16
Ferrari Maranello	16
Ford Thunderbird	23
Honda Insight	66
Lamborghini Gallardo	15
Lamborghini Murcielago	13
Lotus Esprit	22
Maserati Spyder	17
Mazda Miata	28
Mercedes-Benz SL500	23
Mercedes-Benz SL600	19
Nissan 350Z	26
Prosche Boxster	29
Prosche Carrera 911	23
Toyota MR@	32

If we consider a stemplot of this data we have:

1	3 5 6 6 7 9
2	0 2 3 3 3 4 5 5 6 8 8 8 9
3	2
4	
5	
6	6

Key: 1 | 3 = 13mpg

We can see the data is unimodal and skewed right, perhaps even significantly skewed right. If we calculate the mean mpg for these sports cars we get $\bar{x} = 24.7$ mpg. Also, we find the median mpg for these sports cars to be $x = 23$ mpg. *(You should verify that you know how to find these values.)* Notice that the mean is <u>larger</u> than the median. This is not a coincidence, it is a result of the fact that our data is skewed right.

Does the data appear to contain an outlier? In looking at our stemplot, the value of 66 mpg appears to be significantly different from the other data values, so we suspect it may be an outlier. Later in this packet we will learn exactly how to determine if a value is "different enough" to be considered an outlier. For now, let's suppose 66 is an outlier.

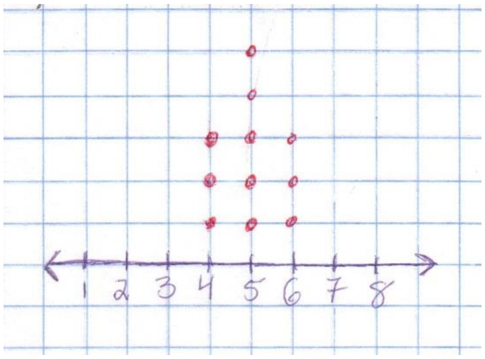
How does this outlier influence our measures of center? Although we <u>never</u> want to ignore outliers, or remove them permanently from a data set, for our purposes we will temporarily remove the 66 from the data. Without the outlier, we calculate the following: $\bar{x} = 22.6$ mpg and $x = 23$ mpg. Notice the mean has changed a great deal when the outlier was removed, but the median has not. In fact, without the outlier in our data set the mean and median are almost equal!

When we are considering data that has skew or outliers, we want to be sure we are accurately measuring the entire data set, and not just the influence of one or a few extreme values. Because of this, we seek to use measures that are ***Resistant***. A measure of center is called ***Resistant*** if it is not subject to the effects of extreme values such as outliers. The median is a resistant measure, whereas the mean is not. As a result, when data has skew the mean is pulled toward the skew. So, to get an accurate picture of center we will use the median for skewed data. However, when a data set is symmetric, we can use either the mean or the median to describe center (most people prefer the mean in this instance).

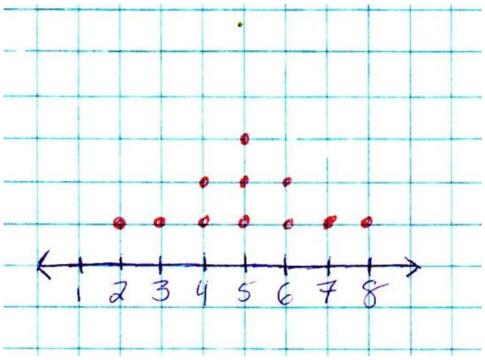
Now, let's turn our attention to the idea of measuring the spread of a set of data.

EXAMPLE: A history teacher has two U.S. History classes. The dotplots below show the number of sources students in each class used to write a History Term Paper.

Class #1:



Class #2:



From the dotplots we can see that both classes have data that is symmetric and unimodal. Additionally, if we calculate the measures of center for each class we have:

Class #1: $\bar{x} = 5$ $x = 5$

Class #2: $\bar{x} = 5$ $x = 5$

Notice the mean and median are equal for both data sets since both are symmetric.

The two sets of class data have the same shape, and the same center. However, clearly the two data sets are not identical. This is why including a measure of spread is important. Shape, center and spread together give us a complete picture of a set of data.

Measures of Spread

1) RANGE – The range of a set of data is the difference between the largest and smallest values in the data set. We do not have a specific symbol to represent the range, largely because it is a very precursory measure of spread that tells only about the extremes, and not very much about the entire data set. As a result, the range is the measure of spread used the least in Statistics.

2) QUARTILES – The quartiles are percentile values that divide the data set into four sections. These measures allow us to compare the spread in the four quarters of the data, not just between the extremes.

The First Quartile (Q_1) – The first quartile is the value larger than 25% of the data. This is the 25th percentile value. Although this can be denoted P_{25}, it is much more commonly denoted Q_1.

The Second Quartile – The second quartile is the value larger than 50% of the data. This is the 50th percentile value. Although this could be denoted P_{50} or Q_2, it really is the median value. So we simply denote the second quartile as x.

The Third Quartile (Q_3) – The third quartile is the value larger than 75% of the data. This is the 75th percentile value. Although this can be denoted P_{75}, it is much more commonly denoted Q_3.

*** There are two other measures of spread to add to our list, but let's work with the quartiles first…*

Consider our Sports Car Highway Fuel Economy example again. The measures of spread for that data are:
range = 66 – 13 = 53 mpg $Q_1 = 18$ $x = 23$ and $Q_3 = 27$.

You should make sure you know how to find these values!

Once you have done that, notice that there is a difference of 5 values between our minimum value and Q_1, a difference of 5 values between Q_1 and x, a difference of 4 values between x and Q_3, and a difference of 39 between Q_3 and the maximum. This suggests that most of the data (the lower 75%) is grouped closed together, but the top 25% of the data is very spread out. As a result, the data is skewed right. Notice, this description does match the picture we see in our stemplot of the data.

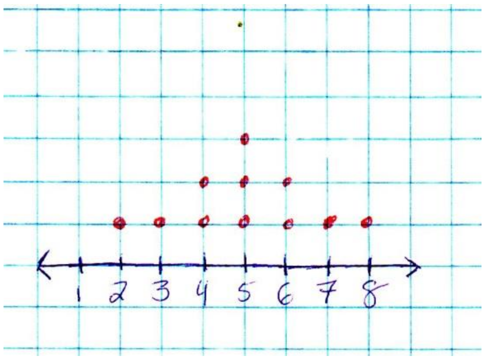
So, the quartiles can tell us a great deal about where the spread lies in a distribution. However, it can be cumbersome to have 3 data values (Q_1, x, and Q_3) to describe the spread of one distribution. This is where the next idea of measuring spread came from:

Could we measure how far each data value is from the center of the distribution? If so, it seems reasonable that this measure would give us a good picture of how spread out the data is. If the *Average Distance from Center* was a small number, the data would be close to the center. But, if the *Average Distance from Center* was a large number, the data would be much farther from center, and therefore more spread out.

↘

To consider how we might measure *Average Distance from Center*, let's consider our History Class data from Class #2 again:

Class #2:



We already know that, for this data, $\bar{x} = 5$. But, how far, on average, are the other data points from the center of 5 sources? To measure this we will calculate:

$$\textit{Average Distance from Center} = a.d.f.c. = \frac{\sum(x-\bar{x})}{n}$$

That is, we will measure how far each data value lies from the center (by calculating $x-\bar{x}$), add up all of those distances, then divide by how many data values we have.

So, for our Class #2 data:

$$a.d.f.c. = \frac{(2-5)+(3-5)+(4-5)+(4-5)+(5-5)+(5-5)+(5-5)+(6-5)+(6-5)+(7-5)+(8-5)}{11}$$

And, if we finish this calculation, we get *a.d.f.c. = 0.*

Clearly this value is not going to help us describe how spread out the data is for Class #2. But why did this happen, and how can we fix it? The problem is, we subtracted the mean from each of our data values. So, some of the distances from center were counted as positive values (for the data values greater than 5), and some were counted as negative values (for the data values less than 5). Also, since each of the data values were used to calculate the mean, *a.d.f.c. = 0* will happen for any data set!

Because of this, we need to do something different. There are a few ways we could fix this, but keeping in mind that we want one equation that will help us measure spread (not two different ways to calculate based on whether the data value is above or below the mean), statisticians decided to square the distances to make sure all of them were counted as positive values. As a result we have:

$$\textit{Average \underline{Squared} Distance from Center} = \frac{\sum(x-\bar{x})^2}{n}$$

Notice for this equation you find the difference of each data value and the mean, square the difference, then add up all of those squared differences. Then, the total is divided by the number of data values you have, *n*.

If we now find the *Average Squared Distance from Center* for our Class #2 data, we have:

$$a.s.d.f.c. = \frac{(2-5)^2+(3-5)^2+(4-5)^2+(4-5)^2+(5-5)^2+(5-5)^2+(5-5)^2+(6-5)^2+(6-5)^2+(7-5)^2+(8-5)^2}{11}$$

And this value is *Average Squared Distance from Center* = 2.73. We now have an actual value that can help us describe how spread out the data is!

However, this is not the end of the story in describing the spread of a set of data. The issue is, what exactly does the *Average Squared Distance from Center* mean? How is it connected to the data we have? This measure is not in the scale of our data because each distance from center was squared. Because of this, we now have the following ideas:

The ***Variance*** of a set of data describes the average squared distance of each observation from the mean. The variance is denoted s^2, and is calculated as:

$$s^2 = \frac{\sum\left(x-\bar{x}\right)^2}{n-1}.$$

Notice the denominator of the variance equation is slightly different from what we originally used to find averaged squared distance. This is done to make or calculation as accurate as possible. Over time Statisticians noticed that variance calculations tended to underestimate the amount of spread in a set of data. So, to slightly increase the value of the variance they reduced the denominator of the formula (because dividing by a smaller number gives a larger answer).

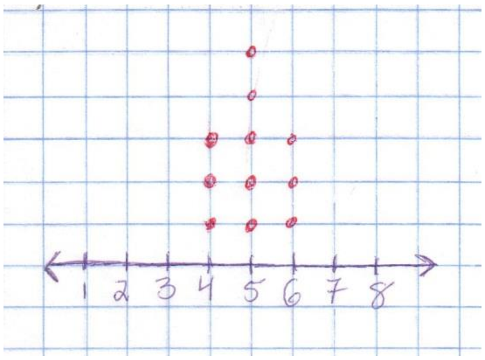
Now, connected with the variance is the idea of ***Standard Deviation***. The standard deviation of a set of data is a measure of the spread of the data in the scale of the data. That is, it describes the average distance from center for the data set, as opposed to the average squared distance from center. To accomplish this, the standard deviation is calculated as:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}.$$

Notice the notation used for standard deviation. Since it is the square root of the variance, it's symbol reflects this idea. Very often, the measure of spread we use is the standard deviation because it is in the scale of our data. However, the variance is still important because we need to calculate it in order to get the standard deviation.

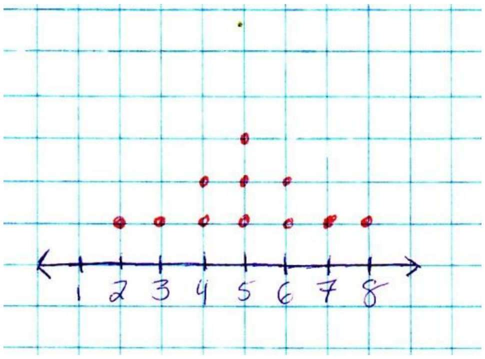
Now, let's again look at our Sources for the History Term Paper example, and consider these new measures of spread.

Class #1



$x = 5$ $\bar{x} = 5$ $s^2 = .6$ $s = .775$

Class #2



$x = 5$ $\bar{x} = 5$ $s^2 = 3$ $s = 1.73$

You should verify that you understand how the variances and standard deviations were calculated. After that, notice that Class #2 has a higher variance and standard deviation than Class #1. We expected this since Class #2 has much more spread in the data set.

Beyond that, we can also use the meaning of standard deviation to interpret the spread of our data. For Class #1, the average distance from center is .775 sources. So, we expect a lot of students to have used between 4.225 and 5.775 sources (or between 4 and 6 sources). For Class #2, the average distance from center is 1.73 sources. So, we expect a lot of students to have used between 3.27 sources and 6.73 sources (or between 3 and 7 sources).

A few final important ideas about standard deviation:

Properties of Standard Deviation

1) Since *s* uses $\bar{x}$ in its calculation, we should use *s* to describe spread when we are using $\bar{x}$ to describe center. If, because of skew, we choose to describe center using x, then we should use the quartiles to describe the spread.
2) If *s* = 0, this means there is no variability or spread in the data. This only occurs when all of the data values are identical. In all other cases, *s* > 0.

3) Like the mean, *s* is not resistant to the effects of outliers. A few extreme values can make *s* very large, even if the rest of the data is close together.

READING SECTION #3 NOTE TAKING:

The following items from this reading must be included in your Statistics notebook:
- Definition, Symbol & Equation for Mean
- Definition & Symbol for Median
- Definition of Mode
- Description of How Skew and Outliers Impact the Measures of Center
- Definition of a Resistant Measure
- Description of When the Mean vs. Median Should Be Used to Describe Center
- Definition of Range
- Definition and Notation of the Quartiles
- Definition, Notation and Equation for Variance
- Definition, Notation and Equation for Standard Deviation
- Properties of Standard Deviation

READING SECTION #4:

In the last lesson we learned about measures of center (mean and median) and measures of spread (quartiles, variance and standard deviation). In the examples we discovered that the mean and standard deviation are not resistant to the effects of extreme skew and outliers, whereas the median and the quartiles are resistant measures. As a result, the median and quartiles are best used to describe non-symmetric data, and the mean and standard deviation are best used to describe symmetric data.

In this lesson we will consider one final method of graphing a set of data. This graph utilizes measures that are resistant, and therefore gives us an unbiased picture of the data distribution. In order to create this graph, we first need the following definition:

Definition: The ***5-Number Summary*** of a set of data describes the shape, center and spread of the data set by displaying the following values: Min Q_1 x Q_3 Max.

EXAMPLE: Suppose a set of data gave us a 5-Number Summary of:
3 14 24 35 92

From the summary we can make the following observations:

- The lower 25% of the data is between 3 and 14, which is 12 possible values.
- The second 25% of the data is between 14 and 24, which is 11 possible values.
- The center of the data is 24.
- The third 25% of the data is between 24 and 35, which is 12 possible values.
- The top 25% of the data is between 35 and 92, which is 58 possible values.
- So, the data is significantly skewed right.

Consider another *EXAMPLE:* The data below is the length, in minutes, of long distance phone calls made by a local business in one day:

3 7 2 14 4 29 3 9 1 20 10 7 2 42 3 5

The 5-Number Summary of this data is: 1, 3, 6, 12, 42 (*You should verify that you understand how these values were arrived at.*) This 5-Number Summary suggests that this data is also skewed right because there is much more spread in the upper 25% of the data than in any other quarter of the data.

Before we work more with this phone call data, we have another concept to consider. One additional calculation commonly used with the 5-Number Summary is the ***Interquartile Range (IQR)***. The Interquartile Range describes the spread of the middle 50% of a data set. It is calculated as: $IQR = Q_3 - Q_1$. This value can be an important measure because it shows us how spread the middle of the data is, and as a result can help us identify how significant any skew might be. It can also be used to determine if a data set contains any outliers.

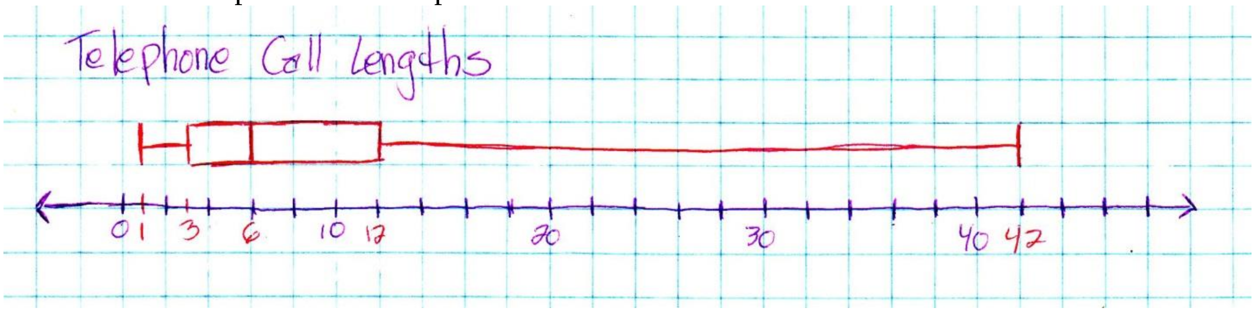
So, in considering our phone call data, the IQR = 12 – 3 = 9. Therefore, the middle 50% of calls made by this business only differ by 9 minutes. Since the top 25% of the data is spread between 12 minutes and 42 minutes, it appears our skew to the right is fairly significant, We may even have an outlier or two on the right end of the data set.

Now that we have an understanding of the 5-Number Summary and the Interquartile Range, we can consider the graph that uses these values, the ***Boxplot***.

Creating a Boxplot

1) Begin by calculating the 5-Number Summary of the data set.
2) The boxplot is a graph created using just one axis. Generally we use a horizontal axis, but the graph can be created vertically as well.
3) Create the axis, and make sure your scale allows you to identify the values in the 5-Number Summary.
4) Draw a vertical line above the axis for each value in the 5-Number Summary.
5) Connect the vertical lines representing Q_1 and Q_3 to create a box. This box represents the middle 50% of the data, and it's length is equal to the IQR. Notice the median line is inside the box.
6) Connect the vertical lines representing the minimum and maximum values to the box using horizontal lines. These lines are called the *whiskers* of the graph.

Below is a boxplot of our Telephone Call data:



Notice the picture the boxplot creates also confirms that our data is skewed right since the right whisker is much longer than the left. Also, notice the median line is not directly in the middle of the box. The more symmetric the data is, the closer the median line will be to the middle of the box, and the closer the whiskers will be to being equal in length. The more skewed a set of data, the farther the median line appears to one side or the other in the box, and the more noticeably the whisker differ in length.

We have mentioned a few times the possibility that this data set contains an outlier, or maybe two. But how do we know for sure? To identify outliers, we think of our boxplot this way: *Consider the length of the box itself. This length is the IQR, and it describes how far apart the middle 50% of the data is. So, when we compare the length of the IQR to the length of a whisker, we consider a value to be significantly far away from the rest of the data if it is 1 ½ times the length of the box away from the middle 50% of the data.*

In other words, we have the following test for identifying outliers:

The Rule for Identifying Outliers

A data value is considered to be an outlier if it lies more than 1 ½ lengths the IQR away from the middle 50% of a data set. That is, a data value is considered an outlier if it is:

Less Than $Q_1 - 1.5(IQR)$

OR

Greater Than $Q_3 + 1.5(IQR)$

If we consider our Telephone Call Example again:

$Q_1 = 3 \qquad Q_3 = 12 \qquad IQR = 9$

Since we believe the outliers are at the higher end of the data set, we calculate:

$Q_3 + 1.5(IQR) = 12 + (1.5)(9) = 25.5$

So, any data value that is greater than 25.5 is an outlier. If we look at our data again, we find that there are two outliers, 29 and 42.

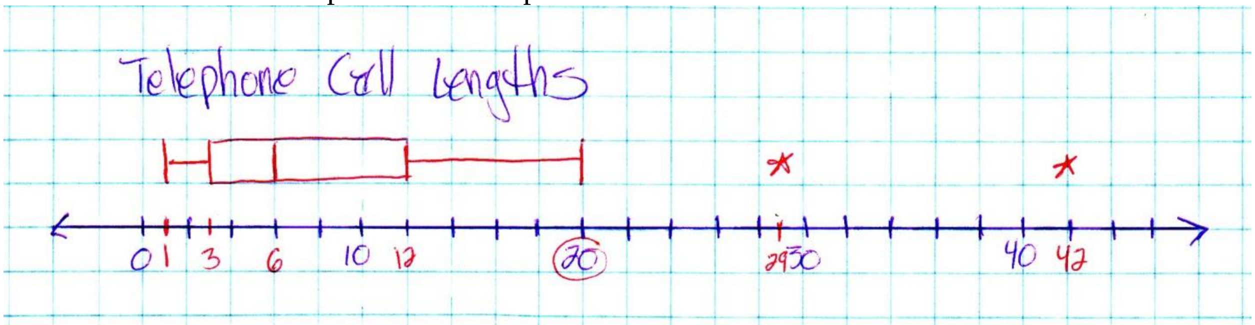
The final idea for this lesson involves how we can identify outliers in our boxplot. If we know they exist, we certainly should identify them! To do this, we create a graph called a ***Modified Boxplot***.

Creating a Modified Boxplot

1) Begin by calculating the 5-Number Summary of the data set.
2) Calculate the threshold and identify any outliers in the data set.
3) Create the axis, and make sure your scale allows you to identify the values in the 5-Number Summary, the outliers, AND *the last data value before the outlier(s) on the side where the outlier(s) exist*.
4) Draw vertical lines above the axis for Q_1, the median and Q_3, then create a box.
5) Draw the vertical line and the whisker for the side of the graph that does not have any outliers.
6) On the side with outliers, draw a vertical line at *the last data value not considered to be an outlier*. This is the line used to draw the whisker.

    ***** It is important to note that the threshold value we calculate to identify outliers is not where the whisker ends. The whisker must end at an actual data value in the data set.***
7) Draw points or stars at the location of any outliers to identify them.

Below is a Modified Boxplot of the Telephone Call Data:



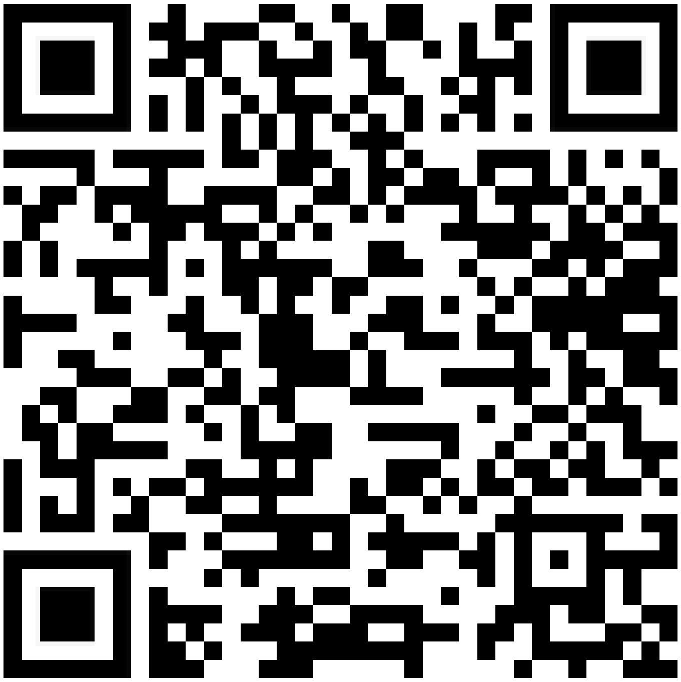
Notice the data would still be described as significantly skewed to the right, and the outliers are still part of our data set. However, the benefit of this graph is allowing you to identify the outliers and determine whether or not they are the sole cause of any skew in the data. In this case, it appears the data is skewed to the right even without the significant influence of the outliers.

READING SECTION #4 NOTE TAKING:

The following items from this reading must be included in your Statistics notebook:

- Definition & Example of the 5-Number Summary
- Definition, Abbreviation & Equation for the Interquartile Range
- Definition, Steps & Example of a Boxplot
- The Rule for Outliers & an Example
- Definition, Steps & Example of a Modified Boxplot

Now that you have completed your Stats Summer Work, finish your assignment by scanning the code below to fill out the Google Form:



***** Make sure you are logged in with your school gmail account to complete the form!***